Finding Time Together

Bano, Sophia; Zhang, Jianguo; McKenna, Stephen

Published in:
2017 IEEE International Conference on Computer Vision Workshop (ICCVW)

DOI:
10.1109/ICCVW.2017.274

Publication date:
2018

Document Version
Peer reviewed version

Finding Time Together: Detection and Classification of Focused Interaction in Egocentric Video

Sophia Bano Jianguo Zhang Stephen J. McKenna
Computer Vision and Image Processing Group, School of Science and Engineering (Computing)
University of Dundee, United Kingdom
{s.bano, j.n.zhang, s.j.z.mckenna}@dundee.ac.uk

Abstract

Focused interaction occurs when co-present individuals, having mutual focus of attention, interact by establishing face-to-face engagement and direct conversation. Face-to-face engagement is often not maintained throughout the entirety of a focused interaction. In this paper, we present an online method for automatic classification of unconstrained egocentric (first-person perspective) videos into segments having no focused interaction, focused interaction when the camera wearer is stationary and focused interaction when the camera wearer is moving. We extract features from both audio and video data streams and perform temporal segmentation by using support vector machines with linear and non-linear kernels. We provide empirical evidence that fusion of visual face track scores, camera motion profile and audio voice activity scores is an effective combination for focused interaction classification.

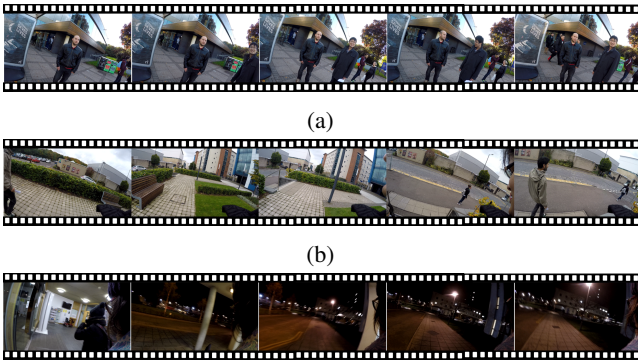
1. Introduction

Recording of daily life experiences from a first-person perspective has become more prevalent with the increasing availability of wearable cameras used in applications such as life-logging, security, sports, ambient assisted living and driving assistance. In recent years, analysis of egocentric video has therefore gained the attention of the computer vision community. Whereas social interaction detection from a third-person perspective has been a well-researched area for some time [4, 10, 25], ego-centric vision-based methods are increasingly addressing the detection and analysis of social interaction from a first-person perspective; methods have been proposed to detect groups of individuals interacting with each other or with the camera wearer [2, 5, 13]. These methods perform off-line processing of short video clips or photo streams mostly captured from constrained perspectives and always containing interacting or non-interacting individuals. However, in reality egocentric videos are unconstrained when used for capturing daily living in long, continuous sequences.

Audio-visual feature fusion has been used for applications such as speaker localisation and event detection in social gatherings using videos captured in highly controlled indoor settings [3, 14], social interaction detection in nursing homes using surveillance-type camera videos [9], and scene change detection in life-logging videos [23]. Although audio signals provide information about social interactions, the fusion of visual and audio cues for detection of social interactions in egocentric video was rarely explored. Furthermore, the effect of integrating global camera motion analysis methods, nowadays used for human activity recognition in egocentric videos [30], with other visual and audio features for social interaction analysis still needs to be researched.

Social interaction occurs when two or more individuals, having mutual focus of attention but not necessarily physically co-present, communicate and interact with one another [27]. Examples include face-to-face verbal conversations, email conversations, and non-verbal (sign language) conversations. Goffman [15] distinguishes between focused and unfocused interactions. *Focused interaction* occurs when two or more co-present individuals, having mutual focus of attention, interact by establishing face-to-face engagement and direct conversation. Note that face-to-face engagement is often not maintained throughout the entirety of a focused interaction. *Unfocused interaction*, on the other hand, occurs when individuals, though co-present, do not establish a direct engagement and conversation [15]. We use the term *conversational partner* for a person who is involved in a focused interaction with the camera wearer. In light of Goffman's theory [15], it is important to highlight that papers describing current egocentric video-based methods [2, 5, 13] that use the term social interaction are actually addressing focused interaction; social interaction is a much broader term.

Figure 1 shows sample frames from three videos in our



(a) (b) (c)

Figure 1: Examples of focused interactions from our *Focused Interaction Dataset*. The frames displayed were sampled at 1 *fps* from the videos. (a) An interaction in which conversational partners are in the field of view of the camera. (b), (c) Interactions in which the conversational partner is not in the field of view as the interactions occurred while walking. (c) An outdoor night-time scenario where the visual cues are weak due to low illumination.

Focused Interaction Dataset[1]. These examples highlight both the variability in egocentric video data in terms of viewpoint, location, and illumination, and the fact that conversational partners are not always in the field of view (e.g., focused interaction while walking). Voice activity and camera motion cues will be especially important in such cases. Voice Activity Detection (VAD) is widely researched in audio signal processing and used for audio conferencing, speech encoding, speech recognition, and speaker recognition [17, 26]. VAD methods detect voice activity (primarily speech) from a noisy audio signal [16, 24, 29]. Video content-based camera motion analysis methods make use of template matching [1] and optical flow [6]. Methods derived from optical flow are widely used nowadays for human activity and action recognition from third person perspective [8, 20] (where a fixed and static camera captures third person activities such that the optical flow is strongly associated with their activity) and first person perspective [30] (where camera wearer activities affect the global camera motion).

Existing methods for social interaction detection in egocentric videos typically assume that people are already present in the field-of-view of the camera and focus on detecting sub-categories of social interaction [13], social groups [5], and presence or absence of social interaction [2] by utilizing visual data only. Fathi *et al.* [13] proposed one of the first methods for detecting different types of social interaction in egocentric video and evaluated it on clips from videos captured at a theme park. They used a multi-label hidden conditional random field model to detect discussion, monologue and dialogue based on estimates of faces' locations and orientations. Alleto *et al.* [5] applied the concept of F-formation [10] for detecting social groups in egocentric video. They designed a pairwise feature vector that describes spatial relationships between two people present based on distances and orientations. A correlation clustering algorithm was applied to merge people into socially related groups. A structural SVM-based method was then used to learn the weight of each component of the correlation clustering vector depending on the social situation. More closely related to our work, Aghaei *et al.* [2] proposed a method for detecting social interaction in low frame rate photo streams. They trained an LSTM-RNN classifier to detect social interaction based on estimates of the distance of an individual from the camera wearer as well as their relative orientation. These social interaction detection methods [2, 5, 13] processed data offline and considered clips of photo streams that always contained people. However, people interacting with the camera wearer may not always be in the field-of-view (e.g. when walking while having a conversation). Moreover, it should be noted that the existing focused interaction methods only consider constrained video segments (clips), where each clip belongs to one specific class, in which the camera wearer is stationary, hence video cues such as face tracking and orientation alone are sufficient [2]. However, a continuously recorded life-logging video has multiple transitions from one class to another.

In this paper, we address the task of identifying temporal segments in continuous egocentric video that correspond to periods of no-focused interaction (no-FI), focused interaction while the camera wearer is stationary (FI-NW) and focused interaction while the camera wearer is walking (FI-W). All such instances of focused interaction should be automatically detected. Often, FI-W is enclosed within FI-NW (e.g., the camera wearer meets a conversational partner, they go for a walk together, and conclude their interaction with a farewell while facing each other). We propose and evaluate a method based on audio and video features to perform the tasks of detection and classification of focused interaction. The main contributions of this paper are as follows.

- We formulate the task of automatic, online classification of focused interactions in continuous, egocentric audio-video data.
- We use spatio-temporal local and global video features and voice-based audio features for classifying focused interaction.
- We propose a temporal segmentation approach based on frame classification and present several variants of it that use Support Vector Machines (SVM) with either

[1] We plan to release the focused interaction dataset after publication

linear or non-linear kernels for classification using various audio-visual feature sets.

- We evaluate the proposed methods on our Focused Interaction dataset, providing empirical evidence that fusion of visual face track scores, camera motion features and voice activity detector scores, and learning using SVMs with non-linear kernels, provides an effective means for classifying focused interactions.

Note that existing face detection and tracking, voice activity detection and global camera motion analysis techniques are adopted in our proposed method as the aim of this work is not to improve these individual techniques but to look into the effect of integrating these techniques to form a robust and online focused interaction system for unconstrained, life-logging, egocentric videos.

The remainder of the paper is organised as follows. The proposed method for focused interaction classification is detailed in Section 2. Section 3 describes our focused interaction dataset and the evaluation protocol. Results of experiments comparing variants of the methods are presented and analysed in Section 4. Finally, Section 5 draws some conclusions.

2. The proposed method

We process audio and video streams independently to extract three distinct features, namely, face track score, camera motion feature vector and voice activity detection score. From the video stream, the face track score is obtained by detecting and tracking faces, and the camera motion feature vector is obtained by computing the histogram of oriented optical flow. From the audio stream, the voice activity detection score is computed by analysing discriminative audio features. These features are fused to form the feature set and SVMs are then trained for the online classification of continuous data streams that contain instances of No-FI, FI-NW and FI-W.

2.1. Face track score

A Histogram of Oriented Gradient (HOG)-based face detector [11, 19] is applied to detect faces in each video frame. Some false and missed face detections are inevitable due to the relatively unconstrained nature of egocentric video. Therefore, we apply a Kanade-Lucas-Tomasi (KLT) point tracker [22, 28] to refine face detection results. Tracking is initialized as soon as a face is detected and continues tracking points of the face in subsequent frames. Track points are updated by taking input from the face detector every tenth frame. The face track is terminated if no face is detected at the same position as that of the tracker or if all points that were tracked are eventually lost.

The KLT tracker returns confidence scores for the point tracks. The neighborhood of the l^{th} point at frame t consists of those pixels in an image patch I_t^l centred on that point. We compute a face tracker score s_t by summing scores of all points tracked on a face, i.e.,

$$s_t = \sum_{l=1}^{L} (1 - \frac{1}{W} ||I_t^l - I_{(t-1)}^l||^2) \quad (1)$$

where W is the number of pixels in a neighbourhood patch. The lower bound for s_t is zero when no faces are tracked while the upper bound depends on the number of points tracked per face (and is certainly no larger than the number of pixels in the face detection box). The track score is high if lots of face points are tracked with confidence. We compute the duration (life) of each track and, in frames in which multiple faces are tracked, we select the one with the longest duration for inclusion in the current feature set. The rationale for this is that short duration tracks often correspond to false detections or to short unfocused interactions (e.g., walking past another person). Moreover, selecting the track with longest duration allows online pruning of tracklets generated through false face detection as these tracklets have comparatively shorter life. Figure 2(a) shows tracker scores obtained from an example video. Representative frames from that video are shown and labelled (i) - (viii). Correct face tracks occur at (ii) and (viii) whereas false face tracks occur at (v) and (vi). As tends to be the case more generally, the true face tracks have greater duration than the false ones.

2.2. Camera motion feature extraction

In the case of first person perspective, the distribution of optical flow produces distinct profiles when the camera is static (e.g., camera wearer standing/sitting) and moving (e.g., camera wearer walking, turning around, going up/downstairs). Histogram of Oriented Optical Flow (HOOF) [8] gives a representation of camera motion at each frame.

Given an input video, we compute dense optical flow using Farneback's method [12]. This results in the flow vectors $v = [x, y]^T$ and their orientations $\theta = \arctan\left(\frac{y}{x}\right)$. HOOF features [8] are computed by binning each flow vector based on its angle with the horizontal axis and weighting it based on its magnitude. The range for the b^{th} bin is defined as

$$2\pi \frac{b-1}{B} \leq \theta < 2\pi \frac{b}{B}, \quad (2)$$

where $1 \leq b < B$ and B is the total number of bins. As a result, we get the normalised histogram $h_t = [h_{t,1}, h_{t,,2}, ..., h_{t,B}]^T$ of HOOF features at each time instant t. h_t is then fused with other features in order to form the feature set.

Figure 2(b) shows HOOF features for a sample video along with the activity description. A focused interaction

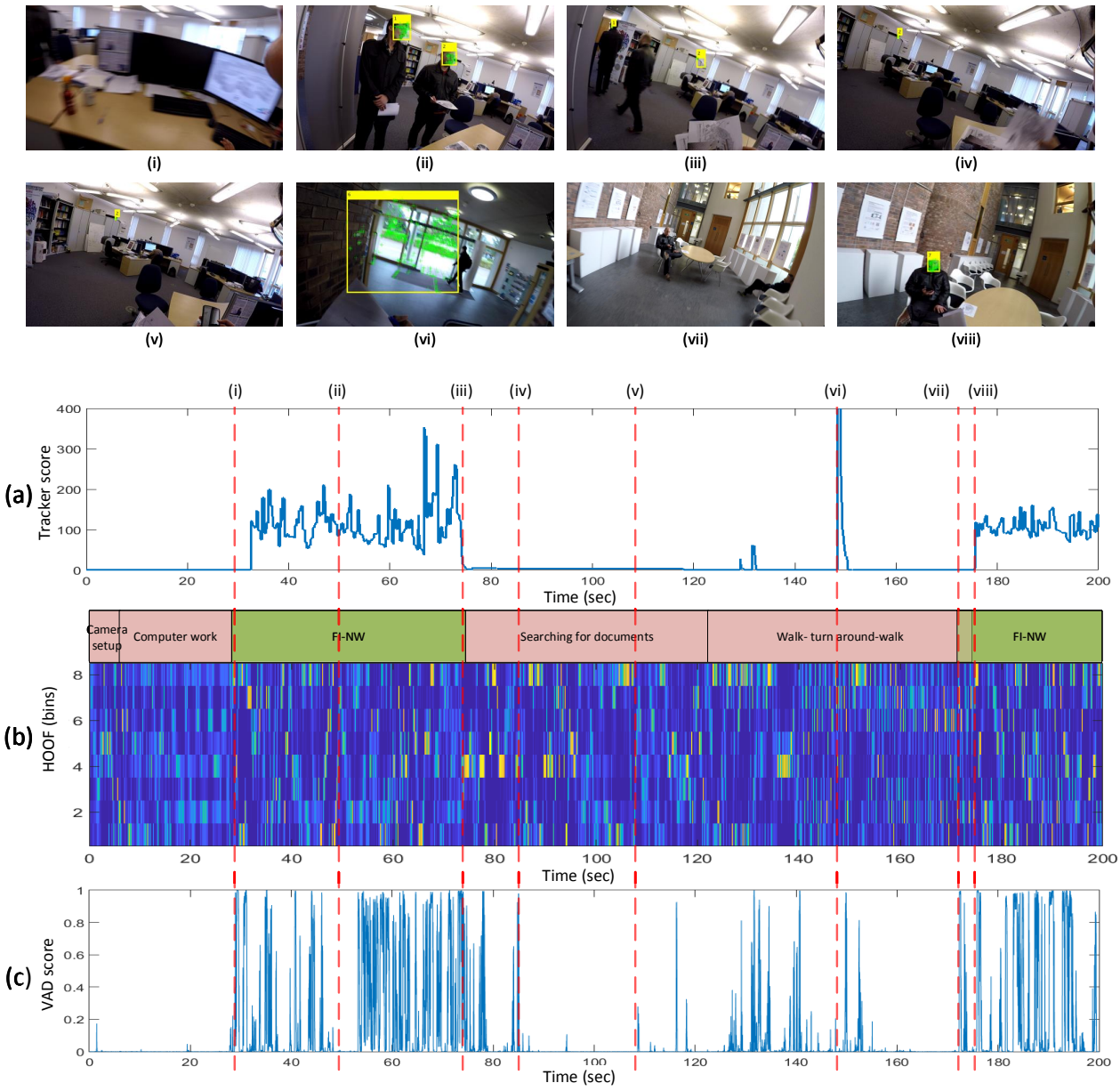


Figure 2: Visualisation of (a) tracker score, (b) Histogram of Oriented Flow (HOOF) features along with activity description, and (c) Voice Activity Detection (VAD) score. Frames (i) to (viii) correspond to the red dashed lines. A focused interaction starts and ends at (i) and (iii), respectively. Another focused interaction starts at (vii). Best viewed in colour.

occurred between (i) and (iii), and another started at (vii). These are both examples of interactions in the FI-NW class. Note that although FI-NW occurs when the camera wearer is stationary, in an active conversation body motion is present; hence variability in HOOF can be observed. Before (i), the camera wearer was engaged in checking emails on a computer; hence the pattern of HOOF features remained consistent. Between (iii) and (vii), the camera wearer searched for some documents and started walking.

2.3. Audio-based feature extraction

We utilize the method and implementation of Segbroeck et al. [29] for voice activity detection (VAD). This method combines four types of discriminative audio features in order to detect voice activity in noisy real-world environments, specifically, spectral shape, spectro-temporal modulations, harmonicity (presence of pitch harmonics) and long-term spectral variability. The resulting VAD

scores range from 0 to 1; a score close to 0 indicates no voice activity while a score close to 1 indicates with high confidence the occurrence of voice activity.

Figure 2(c) shows the estimated VAD scores for a sample video. At (i) a focused interaction begins; although there is no face present in the field-of-view of the camera at this point, voice activity is detected as can be observed in (c). Another focused interaction begins at (vii) but due to motion blur and distance of the participant from the camera, no face is detected until (viii). However, voice activity is detected from the start of this focused interaction. At other times ((iv), (v), (vi)), voice activity is falsely detected albeit with relatively low scores.

2.4. Audio-visual feature fusion

Video and audio features are obtained at different sampling rates; visual features are updated once every video frame, i.e. at $25Hz$, whereas VAD scores are computed every $10\ ms$, i.e. at $100Hz$, given an input audio stream with sampling rate of $8000\ Hz$ (the default setting proposed by [29]). In order to fuse these features we resample the audio features, specifically we average four consecutive VAD scores, with a step size of four, to get the score at the same rate as that of the video features.

The tracker score s_t, HOOF h_t and VAD score v_t are normalised to have zero-mean and unit variance based on estimates of their mean and variance obtained from training data, resulting in $\hat{s}_t$, $\hat{h}_t$ and $\hat{v}_t$. In each frame, the three extracted features are concatenated to form the feature vector $f_t = [\hat{s}_t, \hat{h}_t, \hat{v}_t]^T$.

2.5. Focused interaction classification

The task to be performed is to sequentially process the input audio-visual data stream in order to identify temporal segments corresponding to periods of No-FI, FI-NW and FI-W. One way to formulate a solution is to classify each frame as belonging to one of the three classes. We train Support Vector Machines (SVM) for classification by using the features extracted from a fixed-length temporal window and the ground-truth label for each window.

We trained SVMs on feature vectors which were the concatenation of the audio and video features extracted from each of M consecutive frames. The goal was to assign to each temporal window of M frames, the classification label for the frame at the middle of the window. Windows were extracted with a shift of H frames so that a classification was obtained every H^{th} frame. LIBSVM [7] was used for training the SVM using either linear or RBF kernel.

3. Experimental Setup

3.1. Dataset

The number of annotated datasets publicly available for research that capture social interactions using ego-centric cameras is limited, in part due to privacy concerns. Some datasets [21] are only partially available, without audio and with anonymized (blurred) faces of people in the field-of-view of the camera. Others captured photostreams at 2 frames per minute of social interaction (without audio) [2] (not yet publicly available) or of multiple people interacting as social groups at 4 different locations [5]. Another dataset captured in a theme park is labelled for three different types of social interaction (dialogue, discussion and monologue) as well as for activities (e.g., walking, waiting, gathering, sitting). However, our everyday scenarios are significantly different from activities performed in a theme park. Therefore, we recorded a *Focused Interaction* dataset that captured various focused and unfocused interactions interspersed naturally with periods of no interaction, in real-world unconstrained scenarios and in varying environmental conditions (e.g. indoor/outdoor, daylight/night).

Our *Focused Interaction Dataset* contains 19 egocentric continuous videos captured, at high resolution (1080p) and at a frame rate of 25*fps*, using a shoulder-mounted GoPro Hero4 camera and a smartphone (for inertial and GPS data), at 18 different locations and with 16 different conversational partners. This makes our dataset useful for other egocentric applications such as scene categorization and person association. A shoulder-mounted camera is preferred over a head-mounted one as it is less obstructive and provides relatively stable video because the camera does not move with the user's head. Our dataset contains 378 minutes of recordings (approximately 560k video frames) annotated into periods of No-FI, FI-NW and FI-W. The dataset is unconstrained in nature as neither the camera wearer nor the conversational partners were given any specific instructions that may restrict their movement and it was captured at several indoor and outdoor locations at different times of the day and night, and in different environmental conditions (e.g. sunny or cloudy, with background noise from nearby people and cars). In total, 240 *mins* (64%) of data contain focused interaction in which conversational partners are in the field-of-view of the camera most of the time and are not walking; their positions and face orientations vary significantly. 50 *mins* (13%) of data contain focused interactions in which the conversational partners are not in the field-of-view of the camera and 88 *mins* (23%) of data do not contain any focused interaction.

3.2. Evaluation protocol

We used seven different feature sets for evaluation, namely, TVM, TV, TM, VM, T, V and M, where T de-

Table 1: Evaluation of various feature sets and SVM kernels when using one-versus-all SVM classification. Values given are pooled over the 6 validation folds. Key: $\mathcal{P}$ - precision; $\mathcal{R}$ - recall; $\mathcal{F}$ - F1-score; AUC - area under curve; T - track score; V: VAD score; M - motion vector; FI - focused interaction; NW - non-walk; W - walk.

		Linear kernel					Non-linear (RBF) kernel					
Feature set	Class	C	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	AUC	C	γ	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	AUC
TVM	No-FI	2^{-9}	93.54	93.84	**93.69**	92.99	2^2	2^{-9}	94.05	94.89	**94.46**	**94.65**
	FI-NW		85.16	91.59	88.25	94.31			90.20	87.86	89.01	**96.46**
	FI-W		91.30	97.04	**94.08**	**89.78**			93.70	97.62	**95.62**	**94.53**
TV	No-FI	2^{-6}	93.65	93.62	93.63	**93.46**	2^2	2^{-7}	93.33	95.06	94.19	92.19
	FI-NW		83.79	93.39	**88.33**	**95.00**			87.61	92.20	**89.85**	94.63
	FI-W		89.59	98.02	93.62	87.42			92.76	96.17	94.44	88.63
TM	No-FI	2^{-9}	83.23	88.47	85.77	82.65	2^2	2^{-9}	90.23	89.74	89.98	90.32
	FI-NW		83.28	92.90	87.82	93.10			86.96	89.06	87.99	95.58
	FI-W		86.94	98.30	92.28	79.31			91.81	97.20	94.43	91.08
VM	No-FI	2^{-9}	90.82	92.00	91.41	88.17	2^2	2^{-9}	91.96	94.38	93.16	91.75
	FI-NW		69.79	59.51	64.24	77.96			86.39	76.742	81.28	91.96
	FI-W		86.81	99.99	92.93	47.56			92.29	97.81	94.97	90.22
T	No-FI	2^0	79.19	81.93	80.54	78.05	2^1	2^{-9}	87.53	80.27	83.74	81.02
	FI-NW		82.33	93.11	87.39	93.68			87.97	90.35	89.15	92.28
	FI-W		86.81	99.90	92.90	35.27			86.81	99.36	92.66	55.43
V	No-FI	2^{-10}	91.08	91.31	91.19	89.22	2^{-6}	2^{-5}	89.07	94.36	91.64	83.19
	FI-NW		67.53	59.65	63.35	76.37			76.96	50.99	61.34	71.82
	FI-W		86.81	100.0	92.94	26.48			86.81	100.0	92.94	28.25
M	No-FI	2^{-9}	76.82	99.98	86.88	48.69	2^2	2^{-9}	84.58	91.96	88.12	80.46
	FI-NW		62.20	28.99	39.54	65.58			82.49	66.86	73.86	86.64
	FI-W		86.81	100.0	92.94	44.98			90.93	97.43	94.07	86.55

notes the face track score, V denotes the VAD score and M denotes the camera motion feature vector. SVM with either linear or RBF kernel in one-versus-all setting was used for the classification using different feature sets. For each feature set, the best performing SVM parameters, C and γ, for linear and RBF kernels were computed using grid search [18] (reported in Table 1) and were then used for the validation. A temporal window size of 50 *frames* (selected empirically) with a shift of 25 *frames* was used for obtaining the training samples, while testing was performed with a shift of 1 *frame*.

Evaluation was performed using six-fold division of our focused interaction dataset. Since the duration of each recording varied, it was not possible to have exactly equal numbers of frames in each fold without arbitrarily breaking videos into smaller parts. Instead the folds were generated to roughly contain 60 *mins* of data.

We use framewise evaluation measures to assess the performance by comparing the predicted labels against the ground-truth labels. For each class, we plot the Receiver Operating Characteristic (ROC) curve using the one-versus-all strategy and compute the Area Under the Curve (AUC). The Precision, $\mathcal{P}$, Recall, $\mathcal{R}$, and F1-score, $\mathcal{F}$, are then reported. For the three-class confusion matrix, the predicted class labels are obtained by assigning a positive label to the class with maximum score among the three classes.

4. Results and Discussion

Table 1 summarises the one-versus-all evaluation results for the different feature sets using either linear or RBF kernel for SVM. The corresponding ROC curves are shown in Fig. 3.

TVM-RBF outperformed other feature sets and linear kernel giving an AUC of 94.65% for No-FI, 96.46% for FI-NW and 94.53% for FI-W, respectively. TVM-RBF gave F1-score of 94.46% for No-FI, 89.01% (slightly lower than F1-score for TV-RBF) for FI-NW and 95.62% for FI-W, respectively. In the case of No-FI, voice activity and face tracks are not present most of the time but different types of camera motion occur, e.g., camera wearer walking, sitting or standing alone. Hence motion feature (M) alone cannot reliably identify this class. Observe the low performance of M in Fig. 3(a), (d). From these figures and Table 1, it can be observed that the performance is comparable when using TVM-Linear and TV-Linear for No-FI, while the use

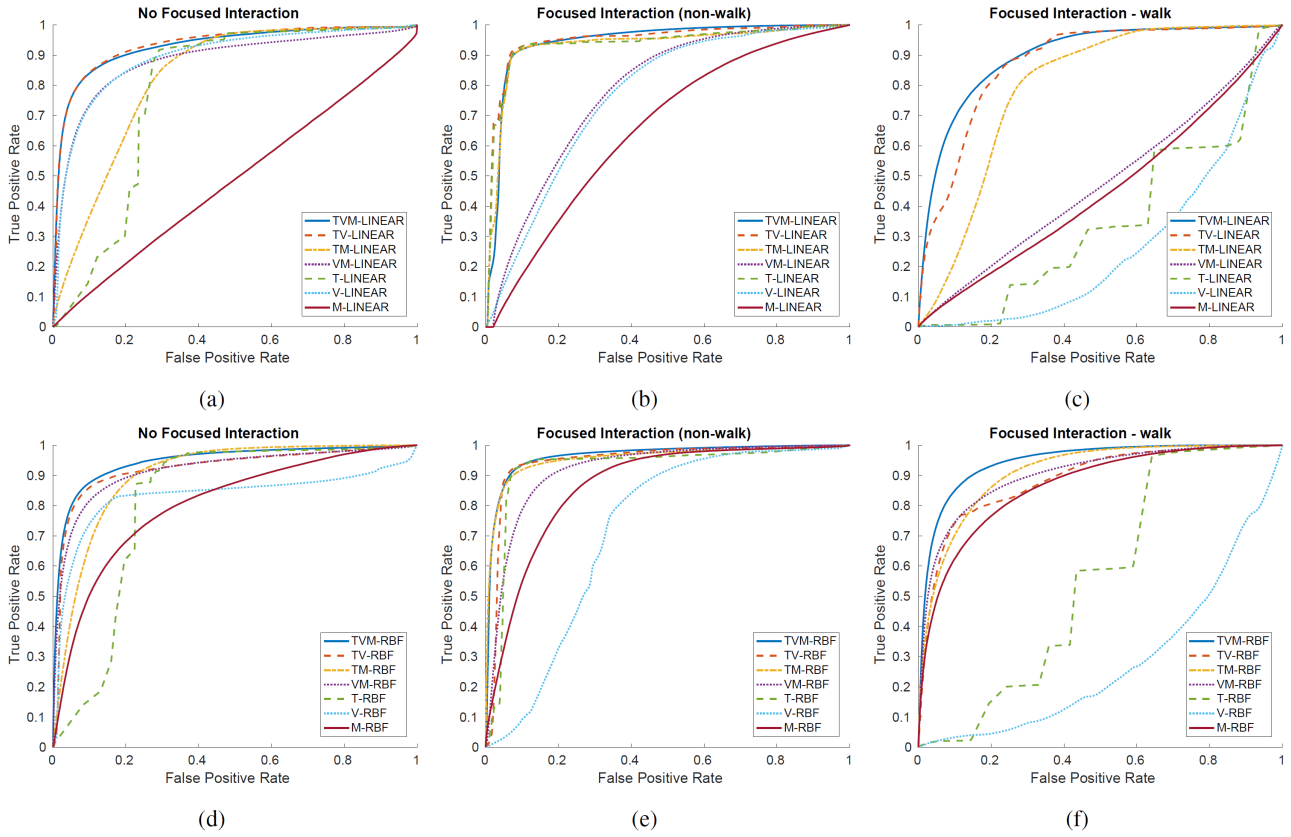


Figure 3: ROC for the various feature sets and SVM with linear (1st row) and RBF (2nd row) kernels. (a) and (d) report the performance of No-FI, (b) and (e) show the results for FW-NW, and (c) and (f) show the results for FI-W, when using linear and RBF kernels. Best viewed in colour.

of RBF kernel helped to improve the performance (AUC increased by 2% for TVM). The instances of FI-NW contain face-to-face interaction, where face tracks are present most of the time. Hence TVM, TV and T with both linear and RBF kernels performed equally well (Fig. 3(b), (e)). M provides a strong cue for discriminating between FI-NW and FI-W. Face tracks are not present in FI-W as the conversational partner is not in the field-of-view of the camera. Moreover, FI-W occurred mostly in outdoor scenarios where the audio signal might get corrupted by background noise (from roadside and passing-by people). Hence, the performance of FI-W is extremely low when using T or V alone. It can be observed from Fig. 3 (comparing the 1st row with the 2nd row) that M has a non-linear relation with the focused interaction classes. This is evident from lower performance of M when using a linear kernel but significantly improved performance when using a non-linear kernel.

The confusion matrices for top performing feature sets are shown in Fig. 4. A confusion matrix is computed by selecting the class with maximum score at each frame in the one-versus-all strategy. TVM-Linear gave accuracy of 62.3% for No-FI, 95.1% for FI-NW and 70.6% for FI-W (Fig, 4(a)). TVM-RBF outperformed by giving accuracy of 73.3% for No-FI, 93.3% for FI-NW and 80.7% for FI-W (Fig, 4(f)). RBF kernel with M, in particular, is useful for classifying FI-W as this interaction occurred while walking. As observed from the ROC curves (Fig. 3), TVM and TV gave comparable results for FI-NW as motion features do not contribute much. It can be observed from the confusion matrices that the overall accuracy when using TV as feature set with linear kernel was 80.9% (Fig. 4(b)) and with RBF kernel was 85.5% (Fig. 4(g)). Similar trend is observed when using TVM as feature set giving an overall accuracy of 82.6% (Fig. 4(a)) with linear kernel and 86.8% (Fig. 4(f)) with RBF kernel, suggesting that the use of non-linear SVM kernel helped in improving the performance.

5. Conclusion

We have presented a method for the automatic online classification of focused interaction in continuous, egocentric videos captured in unconstrained everyday scenarios.

Figure 4: Confusion matrices obtained using different feature sets and SVMs with linear (upper row) and non-linear (lower row) kernels.

We processed both audio and video data streams to obtain audio-visual feature sets. In particular, fusion of face track scores and camera motion profile extracted from visual data with voice activity detection scores from audio data proved to be effective. We performed temporal segmentation of focused interactions via classification using SVMs with different kernels. We evaluated variants of the methods, including single and multimodal feature sets. The use of camera motion profile along with face track and voice activity detection scores and SVM with non-linear kernel were in particular useful for discriminating no focused interaction and focused interaction while walking. Face track and voice activity detection scores were significant for discriminating face-to-face focused interaction for which SVM with non-linear kernel and camera motion profile did not give any further improvement.

In future, we plan to extend this work to identify conversational partners even when they are not in the field-of-view of the camera (e.g., focused interaction while walking) to enhance assistive technology for non-speaking people wearing an egocentric camera.

Acknowledgements. This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant EP/N014278/1: ACE-LP: Augmenting Communication using Environmental Data to drive Language Prediction. The authors are grateful to Annalu Waller (University of Dundee), the ACE-LP team and CVIP members (University of Dundee) for useful discussions and assistance with dataset collection.

References

[1] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp. Camera motion-based analysis of user generated video. *IEEE Transactions on Multimedia*, 12(1):28–41, 2010.

[2] M. Aghaei, M. Dimiccoli, and P. Radeva. With whom do I interact? Detecting social interactions in egocentric photo-streams. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, pages 2959–2964. IEEE, 2016.

[3] X. Alameda-Pineda, V. Khalidov, R. Horaud, and F. Forbes. Finding audio-visual events in informal social gatherings. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 247–254. ACM, 2011.

[4] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of the 23rd International Conference on Multimedia*, pages 5–14. ACM, 2015.

[5] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 580–585. IEEE, 2014.

[6] J. Almeida, R. Minetto, T. Almeida, R. da S. Torres, and N. Leite. Robust estimation of camera motion using optical flow models. *Advances in Visual Computing*, pages 435–446, 2009.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[8] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on

nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1932–1939. IEEE, 2009.

[9] D. Chen, J. Yang, R. Malkin, and H. D. Wactlar. Detecting social interactions of the elderly in a nursing home environment. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):6, 2007.

[10] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *Proceedings of the British Machine Vision Conference*, volume 2, page 4, 2011.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[12] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, LNCS 2749, pages 363–370, 2003.

[13] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012.

[14] I. D. Gebru, X. Alameda-Pineda, R. Horaud, and F. Forbes. Audio-visual speaker localization via weighted clustering. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014.

[15] E. Goffman. *Encounters: Two studies in the sociology of interaction.* Bobbs-Merrill, 1961.

[16] S. Graf, T. Herbig, M. Buck, and G. Schmidt. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(1):91, 2015.

[17] S. Hizlisoy and Z. Tufekci. Noise robust speech recognition using parallel model compensation and voice activity detection methods. In *Proceedings of the 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, pages 1–4. IEEE, 2016.

[18] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[19] D. E. King. Dlib-ML: A machine learning toolkit. *Journal of Machine Learning Research*, 10(July):1755–1758, 2009.

[20] S. S. Kumar and M. John. Human activity recognition using optical flow based feature set. In *Proceedings of the IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–5. IEEE, 2016.

[21] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353. IEEE, 2012.

[22] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674 – 679, 1981.

[23] K. Mahkonen, J.-K. Kämäräinen, and T. Virtanen. Lifelog scene change detection using cascades of audio and video detectors. In *Proceedings of the Asian Conference on Computer Vision*, pages 434–444. Springer, 2014.

[24] M.-W. Mak and H.-B. Yu. A study of voice activity detection techniques for NIST speaker recognition evaluations. *Computer Speech & Language*, 28(1):295–313, 2014.

[25] M. J. Marín-Jiménez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014.

[26] J. Ramirez, J. M. Górriz, and J. C. Segura. *Voice activity detection. fundamentals and speech recognition system robustness*. INTECH Open Access Publisher NewYork, 2007.

[27] R. J. Rummel. *Understanding Conflict and War: Vol. 2: The Conflict Helix: Chapter 9: Social Behavior and Interaction*. Beverly Hills, California: Sage Publications, 1976.

[28] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, School of Computer Science, Carnegie Mellon University, 1991.

[29] M. Van Segbroeck, A. Tsiartas, and S. Narayanan. A robust frontend for vad: Exploiting contextual, discriminative and spectral cues of human voice. In *INTERSPEECH*, pages 704–708, 2013.

[30] K. Zhan, F. Ramos, and S. Faux. Activity recognition from a wearable camera. In *Proceedings of the 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 365–370. IEEE, 2012.